CAICT 中国信通院

人工智能数据安全
白皮书
(2019 年)

中国信息通信研究院
安全研究所
2019年8月

版权声明

前　言

人工智能作为引领新一轮科技革命和产业变革的战略性技术，已成为世界主要国家谋求新一轮国家科技竞争主导权的关键领域。随着政府人工智能战略布局的落地实施，全球人工智能发展正进入技术创新迭代持续加速和融合应用拓展深化的新阶段，深刻改变着国家政治、经济、社会、国防等领域的运行模式，对人类生产生活带来翻天覆地的变化。

数据作为驱动本轮人工智能浪潮全面兴起的三大基础要素之一，数据安全风险已成为影响人工智能安全发展的关键因素。与此同时，人工智能应用也给数据安全带来严峻挑战，如何应对人工智能场景下的数据安全风险日渐成为国际人工智能治理的重要议题。部分国家已率先探索人工智能数据安全风险的前瞻研究和主动预防，并积极推动人工智能在数据安全领域应用，力求实现人工智能与数据安全的良性互动发展。

本白皮书从人工智能数据安全的内涵出发，首次提出人工智能数据安全的体系架构，在系统梳理人工智能数据安全风险和安全应用情况的基础上，总结了国内外人工智能数据安全治理现状，研究提出了我国人工智能数据安全治理建议。

目 录

一、 人工智能数据安全概述

（一） 人工智能安全

当前，由人工智能引领的新一轮科技革命和产业变革方兴未艾，正在对经济发展、社会进步、国家治理等方面产生重大而深远的影响。世界主要国家和全球产业界高度重视并积极布局，人工智能迎来新的发展浪潮。然而，技术进步往往是一把“双刃剑”，本项目组在《人工智能安全白皮书（2018 年）》中提出人工智能因其技术的局限性和应用的广泛性，给网络安全、数据安全、算法安全和信息安全带来风险，并对国家政治、军事和社会安全带来诸多挑战。与此同时，人工智能因其突出的数据分析、知识提取、自主学习、智能决策等能力，可在网络防护、数据管理、信息审查、智能安防、金融风控、舆情监测等网络信息安全领域和社会公共安全领域有许多创新性应用。为有效管控人工智能安全风险并积极促进人工智能技术在安全领域应用，可从法规政策、标准规范、技术手段、安全评估、人才队伍、可控生态等方面构建人工智能安全管理体系。

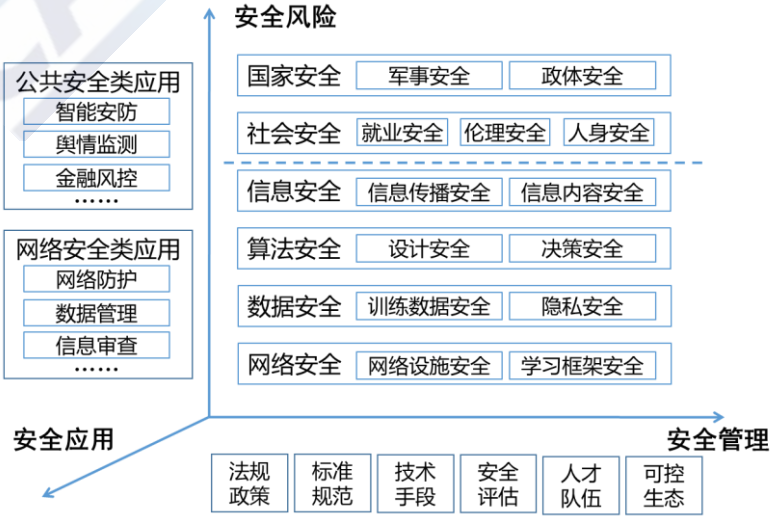

图 1 人工智能安全体系架构图

（二）人工智能数据安全内涵

1、人工智能与数据

人工智能与数据相辅相成、互促发展。一方面，海量优质数据助力人工智能发展。现阶段，以深度学习为代表的人工智能算法设计与优化需要以海量优质数据为驱动。谷歌研究提出，随着训练数据数量级的增加，相同机器视觉算法模型的性能呈线性上升。牛津大学国际发展研究中心将大数据质量和可用性作为评价政府人工智能准备指数的重要考察项[1]。美国欧亚集团咨询公司将数据数量和质量视为衡量人工智能发展潜力的重要评价指标[2]。**另一方面，人工智能显著提升数据收集管理能力和数据挖掘利用水平。**人工智能在人们日常生活和企业生产经营中大规模应用，获取、收集和分析更多用户和企业数据，促进人工智能语义分析、内容理解、模式识别等方面技术能力进一步优化，更好地实现对收集的海量数据进行快速分析和分类管理。而且，人工智能对看似毫不相关的海量数据进行深度挖掘分析，发现经济社会运行规律、用户心理和行为特征等新知识。基于新知识，人工智能进一步提升对未来的预测和对现实问题的实时决策能力，提升数据资源利用价值，优化企业经营决策、创新经济发展方式、完善社会治理体系。

2、人工智能数据安全

数据安全是人工智能安全的关键。数据成为本轮人工智能浪潮兴起发展的关键要素。人工智能算法设计与优化需要以海量优质数据资

[1] 《2019 年政府人工智能准备指数》
[2] 《中国拥抱 AI》

源为基础。数据质量和安全直接影响人工智能系统算法模型的准确性，进而威胁人工智能应用安全。与此同时，人工智能显著提升数据收集管理能力和数据价值挖掘利用水平。人工智能这些能力一旦被不当或恶意利用，不仅威胁个人隐私和企业资产安全，甚至影响社会稳定和国家安全。而且，人工智能、大数据与实体经济不断深度融合，成为推动数字经济和智能社会发展的关键要素。人工智能大规模应用间接促使数据权属问题、数据违规跨境等数据治理挑战进一步加剧。

人工智能为数据安全治理带来新机遇。人工智能驱动数据安全治理加速向自动化、智能化、高效化、精准化方向演进。人工智能自动学习和自主决策能力可有效缓解现有数据安全技术手段对专业人员分析判断的高度依赖，实现对动态变化数据安全风险的自动和智能监测防护。人工智能卓越的海量数据处理能力可有效弥补现有数据安全技术手段数据处理能力不足的缺陷，实现对大规模数据资产和数据活动的高效、精准管理和保护。人工智能赋能数据安全治理，助力数据大规模安全应用，将有力推动经济社会数字化转型升级。

基于以上分析，项目组认为，人工智能数据安全内涵包含：一是应对人工智能自身面临和应用导致及加剧的数据安全风险与治理挑战；二是促进人工智能在数据安全领域中的应用；三是构建人工智能数据安全治理体系，保障人工智能安全稳步发展。

（三）人工智能数据安全体系架构

图 2 人工智能数据安全体系架构图

基于对人工智能数据安全内涵分析，项目组提出覆盖人工智能数据安全风险、人工智能数据安全应用、人工智能数据安全治理三个维度的人工智能数据安全体系架构。其中，**人工智能数据安全风险**是人工智能数据安全治理的起因，包含人工智能自身面临的数据安全风险，人工智能应用导致的数据安全风险，人工智能应用加剧的数据治理挑战。本白皮书重点分析人工智能相关特有数据安全风险与治理挑战。**人工智能数据安全应用**是人工智能技术用于数据安全治理，包含人工智能技术在精准化数据安全策略制定、自动化数据资产安全管理、智能化数据活动安全保护以及高效化数据安全事件管理方面的应用。**人工智能数据安全治理**是应对人工智能数据安全风险和促进人工智

能数据安全应用的体系化方案，包含国家战略、伦理规范、法律法规、监管政策、标准规范、技术手段、人才队伍等方面。

二、 人工智能数据安全风险

（一）人工智能自身面临的数据安全风险

训练数据污染可导致人工智能决策错误。数据投毒通过在训练数据里加入伪装数据、恶意样本等破坏数据的完整性，进而导致训练的算法模型决策出现偏差。数据投毒主要有两种攻击方式：**一种是**采用模型偏斜方式，主要攻击目标是训练数据样本，通过污染训练数据达到改变分类器分类边界的目的。例如，模型偏斜污染训练数据可欺骗分类器将特定的恶意二进制文件标记为良性。**另外一种是**采用反馈误导方式，主要攻击目标是人工智能的学习模型本身，利用模型的用户反馈机制发起攻击，直接向模型“注入”伪装的数据或信息，误导人工智能做出错误判断。随着人工智能与实体经济深度融合，医疗、交通、金融等行业训练数据集建设需求迫切，这就为恶意、伪造数据的注入提供了机会，使得从训练样本环节发动网络攻击成为最直接有效的方法，潜在危害巨大。在自动驾驶领域，数据投毒可导致车辆违反交通规则甚至造成交通事故；在军事领域，通过信息伪装的方式可诱导自主性武器启动或攻击，从而带来毁灭性风险。

运行阶段的数据异常可导致智能系统运行错误。一是人为构造对抗样本攻击，导致智能系统产生错误的决策结果。人工智能算法模型主要反映了数据关联性和特征统计，而没有真正获取数据因果关系。针对算法模型这一缺陷，对抗样本通过对数据输入样例添加难以察觉

的扰动，使算法模型以高置信度给出一个错误的输出。对抗样本攻击可实现逃避检测，例如在生物特征识别应用场景中，对抗样本攻击可欺骗基于人工智能技术的身份鉴别、活体检测系统。2019 年 4 月，比利时鲁汶大学研究人员发现，借助一张设计的打印图案就可以避开人工智能视频监控系统。**二是**动态环境的非常规输入可导致智能系统运行错误。人工智能决策严重依赖训练数据特征分布性和完备性，人工标记数据覆盖不全、训练数据与测试数据同质化等原因常常导致人工智能算法泛化能力差，智能系统在动态环境实际使用中决策可能出现错误。特斯拉汽车自动驾驶系统曾因无法识别蓝天背景下的白色货车，致使发生致命交通事故。

模型窃取攻击可对算法模型的数据进行逆向还原。人工智能算法模型的训练过程依托训练数据，并且在运行过程中会进一步采集数据进行模型优化，相关数据可能涉及到隐私或敏感信息，所以算法模型的机密性非常重要。但是，算法模型在部署应用中需要将公共访问接口发布给用户使用，攻击者可通过公共访问接口对算法模型进行黑盒访问，依据输入信息和输出信息映射关系，在没有算法模型任何先验知识（训练数据、模型参数等）情况下，构造出与目标模型相似度非常高的模型，实现对算法模型的窃取，进而还原出模型训练和运行过程中的数据以及相关隐私信息。新加坡国立大学 Reza Shokri 等针对机器学习模型的隐私泄露问题，提出了一种成员推理攻击，在对模型参数和结构知之甚少的情况下，可以推断某一样本是否在模型的训练

数据集中[3]。

开源学习框架存在安全风险，可导致人工智能系统数据泄露。人工智能开源学习框架实现了基础算法的模块化封装，可以让应用开发人员无需关注底层实现细节，大大提高了人工智能应用的开发效率。谷歌、微软、亚马逊、脸书等企业都发布了自己的人工智能学习框架，在全球得到广泛应用。但是，人工智能开源学习框架集成了大量的第三方软件包和依赖库资源，相关组件缺乏严格的测试管理和安全认证，存在未知安全漏洞。近年来，360、腾讯等企业安全团队曾多次发现 TensorFlow、Caffe、Torch 等深度学习框架及其依赖库的安全漏洞，攻击者可利用相关漏洞篡改或窃取人工智能系统数据。

（二）人工智能应用导致的数据安全风险

人工智能应用可导致个人数据过度采集，加剧隐私泄露风险。随着各类智能设备（如智能手环、智能音箱）和智能系统（如生物特征识别系统、智能医疗系统）的应用普及，人工智能设备和系统对个人信息采集更加直接与全面。相较于互联网对用户上网习惯、消费记录等信息采集，人工智能应用可采集用户人脸、指纹、声纹、虹膜、心跳、基因等具有强个人属性的生物特征信息。这些信息具有唯一性和不变性，一旦被泄露或者滥用会对公民权益将造成严重影响。2018 年 8 月，腾讯安全团队发现亚马逊智能音箱后门，可实现远程窃听并录音。2019 年 2 月，我国人脸识别公司深网视界曝出数据泄露事件，超过 250 万人数据、680 万条记录被泄露，其中包括身份证信息、人

[3] Reza Shokri, Marco Stronati, Congzheng Song, et al.Membership Inference Attacks Against Machine Learning Models

脸识别图像及 GPS 位置记录等。鉴于对个人隐私获取的担忧，智能安防的应用在欧美国家存在较大争议，2019 年 7 月，继旧金山之后，萨默维尔市成为美国第二个禁止人脸识别的城市。

人工智能放大数据偏见歧视影响，威胁社会公平正义。当前，人工智能技术已应用于智慧政务、智慧金融等领域，成为社会治理的重要辅助手段。但是，人工智能训练数据在分布性上往往存在偏差，隐藏特定的社会价值倾向，甚至是社会偏见。例如，海量互联网数据更多体现我国经济发达地区、青壮年网民特征，而对边远地区以及老幼贫弱人群的特征无法有效覆盖。人工智能系统如果受到训练数据潜在的社会偏见或歧视影响，其决策结果势必威胁人类社会的公平正义。在社会招聘领域，美国 Kronos 公司的人工智能雇佣辅助系统让少数族裔、女性或者有心理疾病史的人更难找到工作；在金融征信领域，科技金融公司 Zest 的人工智能信用评估平台 ZAML，采集分析用户网络行为来判定用户的信用值，曾经错误判定不能熟练使用英语的移民群体存在信用问题。

人工智能技术的数据深度挖掘分析加剧数据资源滥用，加大社会治理和国家安全挑战。通过获取用户的地理位置、消费偏好、行为模式等碎片化数据，再利用人工智能技术进行深度挖掘分析，能够预测用户的喜好和习惯，进而对用户进行分类，可实现更加精准的信息推送。基于数据分析的智能推荐可带来用户便利、企业盈利和社会福利，但是也加剧了数据滥用问题。**一是**在社会消费领域，可带来差异化定价。“大数据杀熟”实现对部分消费者的过高定价，甚至进行恶意欺

诈或误导性宣传，导致消费者的知情权、公平交易权等权利受损。2018 年，我国滴滴、携程等均爆出类似事件，根据用户特征实现对不同客户的区别定价，社会负面影响巨大。**二是**在信息传播领域，可引发“信息茧房”效应。人们更多接收满足自己偏好的信息和内容，限于对世界的片面认知，导致社会不同群体的认知鸿沟拉大，个人意志的自由选择受到影响，甚至威胁到社会稳定和国家安全。2018 年曝光的“Facebook 数据泄露”事件中，美国剑桥分析公司利用广告定向、行为分析等智能算法，推送虚假政治广告，进而形成对选民意识形态和政治观点的干预诱导，影响美国大选、英国脱欧等政治事件走向。基于人工智能技术的数据分析与滥用，给数字社会治理和国家安全等带来严峻安全挑战。

人工智能技术可提升网络攻击的智能化水平，进而实施数据智能窃取。一是可用来自动锁定目标，进行数据勒索攻击。人工智能技术可通过对特征库学习自动查找系统漏洞和识别关键目标，提高攻击效率。英国网络安全公司 Darktrace 分析显示，集成人工智能技术的勒索软件可自动瞄准更具吸引力的目标，劫持工业设备、医疗仪器等相关运行数据勒索赎金，受害者为使系统和设备重新上线运行而被迫支付赎金。**二是**自动生成大量虚假威胁情报，对分析系统实施攻击。人工智能通过使用机器学习、数据挖掘和自然语言处理等技术处理安全大数据，能够辅助自动化地生产威胁情报，攻击者也可利用相关技术生成大量错误情报以混淆判断。美国 McAfee 公司指出，“提高噪声基底(noise floor)”技术可对特定环境进行情报轰炸，给威胁情报

分析系统的判断模型制造大量的主动错误信息，造成威胁情报过载，迫使系统重新校准以过滤掉假警报，通过这一过程，攻击者可了解防御逻辑并伺机发起真正的攻击，进而窃取系统数据。**三是**自动识别图像验证码，窃取系统数据。图像验证码是一种防止机器人账户滥用网站或服务的常用验证措施，通过解决视觉难题来验证人类用户，以有效区分拦截恶意程序，保护系统数据安全。但是，人工智能技术已实现对验证码的有效破解。美国 Vicarious 公司开发的基于概率生成模型的验证码识别算法，在标准的 reCAPTCHA 测试中，可成功解开三分之二的验证问题[4]。2017 年，我国浙江省破获了全国第一例人工智能犯罪，案件中黑客利用人工智能识别图片验证码的正确率高达 95%以上，在此平台被打掉前的 3 个月已经提供验证码识别服务 259 亿次。

基于人工智能技术的数据深度伪造将威胁网络安全、社会安全和国家安全。人工智能可利用收集的训练数据进行特征学习，生成逼真的虚假信息内容。特别是近年来基于生成对抗网络（GAN）的“DeepFakes”（深度伪造）技术应用，使得“换脸”虚假视频的制作门槛不断降低，大量深度伪造数据内容开始涌现。我国也出现了徐锦江版“海王”，杨幂版“黄蓉”等逼真虚假视频。目前，深度伪造 2.0 概念已被提出，相比于之前的换脸，深度伪造 2.0 可模仿人的行为举止、声音和习惯动作，更难以区分真假。2019 年 6 月，Facebook 一段扎克伯格的假视频传播迅速，视频里的人从长相、声音、穿衣、手势以及说话时的动作神情都与真人无异。深度伪造数据内容的大量生

[4] Dileep George*, Wolfgang Lehrach,et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs

成和传播，将给网络安全、社会安全和国家安全带来严重风险。**一是**降低生物特征识别技术可信度，提升网络攻击能力。基于图像特征的人脸识别技术和基于声纹的语音识别技术均属于典型的生物特征识别技术，在非接触式身份认证、大流量或自动化安全检测等领域已开展规模化应用。但目前识别伪造音视频存在技术难度，降低了生物特征识别技术的可信度，给网络攻击提供了新手段。**二是**造成人际间的信任危机，威胁伦理和社会安全。随着换脸换声技术的不断进化，伪造图片和音视频的成本会不断降低，各种恶意伪造的图片和音视频信息将大量涌现，会侵犯公民肖像权等个人权益，甚至用于敲诈勒索、伪造罪证等不法活动，从而造成社会信任危机，对伦理道德和社会稳定构成严重威胁。**三是**通过制作虚假新闻影响政治舆论，进而威胁国家安全。国内外恶意势力可利用基于人工智能的换脸换声技术伪造政治领袖和公众人物的新闻视频，普通民众根本无法辨别真假，此类虚假视频内容的大量扩散与传播，可对社会舆论生态造成恶劣影响，引发民众骚乱甚至国内动乱，威胁国家安全。2019 年 6 月，由于担心深度伪造对 2020 年美国大选的灾难性影响，美国众议院已经开始考虑修订现行法案，在立法层面打击相关行为。

（三）人工智能应用加剧的数据治理挑战

人工智能提升数据资源价值，数据权属问题更为突出。一是个人层面，数据权属体现为公民的数据权利，个人隐私保护面临挑战。用户个人隐私信息含金量高，是人工智能技术与产业发展的重要驱动。相关机构在利用用户数据追求自身利益时往往忽视用户个人隐私权

益。近年来，个人隐私泄露重大事件连续发生，顺丰快递、华住酒店、万豪酒店等均出现数亿用户信息泄露事件。另外，互联网用户在使用社交平台、网络直播、在线游戏等应用的过程中，会产生海量社交关系数据和用户行为数据等，这类数据在权利归属上存在争议，但已成为人工智能企业进行算法设计和产品研发的重要支撑。**二是行业层面，**数据权属体现为企业的数据产权，数据垄断损害行业整体发展。人工智能技术使数据经济价值越发凸显，数据已成为企业的核心资产，相关企业积极储备数据资源，并阻止竞争对手获得数据，力图垄断数据资源来最大化企业利益。我国曾爆发华为与腾讯、顺丰与菜鸟之间的数据纠纷事件。数据产权之争将加剧数据垄断。一方面，科技巨头依托网络覆盖和用户规模，加强数据汇聚；另一方面，人工智能中小企业获取数据的渠道受限，数据资源匮乏。企业在数据产权没有被广泛认可，以及数据流动环节存在安全风险的前提下，无论是从维护自身利益角度还是从遵守法律法规角度，都不愿将自身数据进行共享，这将导致初创企业和研究机构在算法设计和优化过程中无数据可用，损害我国人工智能行业整体发展。

人工智能凸显数据的战略地位，数据违规跨境冲击国家安全。当前，世界主要国家都制定了人工智能发展战略，对数据的依赖快速上升，数据作为国家基础性战略资源的地位更加突出。为快速积累数据，科技企业通过向消费者提供特定领域免费应用、使用政府公开数据以及进行产业上下游数据协同等方式获取尽可能多数据。以脸书、谷歌为代表的美国科技巨头，依托其庞大用户规模和强大数据抓取工具，

在全球范围内进行数据收集，强化数据资源优势，推进自身人工智能发展，加剧数据违规跨境流动风险。与此同时，2018 年 3 月，美国发布《澄清境外数据的合法使用法案》(CLOUD 法案)，为美国执法机构访问在美国境内运营的企业存储在海外的用户数据提供明确授权，促使数据管辖权和跨境流动争议进一步加大，威胁我国网络主权和国家安全。

三、 人工智能数据安全应用

（一） 人工智能与数据安全治理

人工智能和数据安全治理互利互补，人工智能技术赋予数据安全治理智慧，数据安全治理为人工智能技术发展提供前驱动力。人工智能技术的发展为数据安全治理提供底层通用技术支撑，取代数据安全治理中大量重复性、长期性、粗略性人类劳动，使数据安全治理向自动化、高效化、精准化、智能化演进。与此同时，数据安全治理工作的开展能提升数据质量，促进数据安全流通和合规使用，为人工智能提供高质量数据集，从而为人工智能技术发展提供前驱动力。具体表现为以下五个方面。

一是人工智能技术可更加准确地理解数据，促进数据安全治理精准化。数据量的丰富为人工智能提供特征广泛的训练数据集，使人工智能模型更加精确。算力的提升使人工智能具备实时数据处理能力，支持在更大范围内及时监测和处理数据，并持续改进样本库，减少样本过少或漏报带来的运算误差。以神经网络为代表的的深度学习技术的发展可以大力提升数据分类分级精准度和数据内容识别准确率。例

如，2012 年神经网络算法只有 5 层，而 2018 年可以做到 1200 多层，在人脸识别领域最高可达一亿分之一的误识率。

二是人工智能技术可取代人类重复性劳动，促进数据安全治理自动化。2018 年李开复在《人工智能》一书中指出，人工智能将在 15 年内具备取代 40-50%岗位的技术能力，主要集中在重复性劳动、有固定台本和对白内容的各种互动、不需与人进行大量面对面交流的工作领域。在数据安全治理领域中，传统的数据特征标注需要大量人力反复筛选和识别，人工智能可以取代人类自动对数据按照内容进行识别和添加标签。在网络安全防护方面，随着网络攻击手段的智能化升级，传统的依赖手动过程以及静态规则和签名的数据传输网络安全保护方法正在失效，人工智能技术可以通过自我学习自动更新安全规则，及时检测出新型网络威胁。

三是人工智能技术直击数据安全治理痛点，促进数据安全治理智能化。数据资产不清晰、数据和知识难以关联、数据安全管理策略更新不及时是数据安全治理中常见问题。与传统数据安全治理相比，人工智能技术可通过精准分级分类自动梳理数据资产，基于统一的管理标准形成元数据，通过智能搜索、关联查询手段，形成数据关联关系图谱，对数据安全风险进行智能评估、量化和预测，辅助形成更合理的安全管理策略。例如，IBM 的大数据安全智能系统实时运用人工智能技术实现了数据的智能高速查询、实时异常检测、自动确定事件根源并开展核查。腾讯的智能大数据治理系统基于基础知识库实现针对不同类型数据的自动感知、智能推荐转换等智能处理功能，人工智能

技术使数据安全治理智能化。

四是人工智能技术可提升系统效率，促进数据安全治理高效化。人工智能可以充分利用自然语言处理、图像识别、语音识别、视频处理等技术弥补传统数据处理耗时长、效率低等弱项，提升系统效率。例如人工智能技术可以对非结构化数据进行高效分析处理，将过去需要几周乃至几个月才能完成的工作缩短到几个小时之内完成，使数据安全治理高效化。华为将机器学习技术用于大数据分析平台，其在中国移动等多个项目的实践表明，数据治理效率提升超过 40%，数据准备周期从月降为小时级，大数据分析应用上线周期从月降到周，同时高效数据治理也提升了数据质量，高质量数据占比提升 40%以上。

五是数据安全治理促进高质量数据集生成，驱动人工智能技术发展。高质量数据集是提升人工智能算法准确性、模型合理性和产品先进性的至关重要的因素，只有当人工智能系统能够获取更为准确、及时、一致的高质量数据，才能提供更高效、更可靠的智能化服务。近年来，随着政府、企业对数据质量管理的重视，数据质量工具市场稳步增长。据 Gartner 发布的 2018 年数据库魔力象限报告显示，2017 年数据质量软件工具市场达到 16.1 亿美元，比 2016 年增长 11.6% 。数据安全治理是提升数据质量的必要途径，是促进人工智能全面发展和应用的基础保障。

（二）人工智能在数据安全治理中的应用

2018 年 5 月，Gartner 发布数据安全治理（Data Security Governance，简称 DSG）框架，提出了从管理层到技术层、从机制体

制到技术工具、全方位覆盖整个组织架构的完整数据安全治理链条。Gartner 指出，直接从数据生命周期环节入手并不合理，需要先确定组织架构，建立管理问责制和决策权，对不同等级的风险制定不同的策略，再利用技术工具对数据全生命周期进行安全风险控制管理，最后对安全风险进行评估并回到第一步重新纠编，形成数据安全治理闭环。2018 年 5 月，中国网信联盟指导下的数据安全治理委员会发布《数据安全治理白皮书》，提出一个通用的数据安全治理框架，将框架分为数据安全治理机制、数据全生命周期管理和数据安全技术部署三个部分。国内外主流数据安全治理框架的思路是相通的，均是以策略机制为入口，以数据全生命周期管理为基础，以技术工具为支撑的多方位治理体系。

本白皮书借鉴国内外主流数据安全治理框架并结合人工智能数据安全应用经验，将人工智能在数据安全治理领域的应用分为数据安全策略制定、数据资产安全管理、数据活动安全保护、数据安全事件管理四个阶段。人工智能技术可应用于数据安全治理的各个阶段，但主要是促进细分领域应用优化升级，距离体系化的智能数据安全治理还有很大差距。如图 3 所示，人工智能数据安全治理细分领域包括数据安全策略、数据分级分类、数据质量管理、数据本体安全保护、数据活动网络安全保护、数据流转行为分析、数据安全风险评估、不良信息治理、互联网反欺诈、打击数据黑产等。

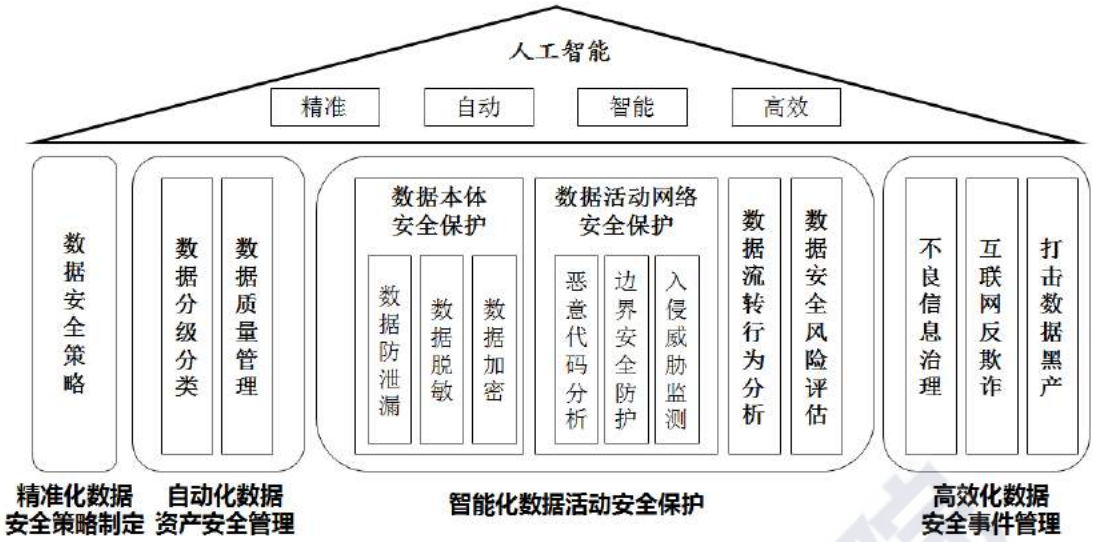

图 3 人工智能在数据安全治理中的应用

1、数据安全策略制定

传统的策略制定过程中用来辅助决策的日志数据和警报数量巨大，决策者难以快速处理，因此传统方式主要依赖人的直觉和经验。人工智能技术具备海量数据采集和分析能力，可根据训练模型进行自我学习并做出相应的判断，使管理更精细、决策更智能，因此智能决策系统应用非常广泛。基于人工智能的决策系统能大大提高数据安全治理策略的时效性和合理性，在数据安全风险管理策略、数据合规性要求、分级保护策略的制定等方面辅助管理者快速、科学、合理地制定策略，为数据安全治理提供智能化的解决方案。例如，2017 年 12 月，百分点集团发布智能政府决策系统 Deep Governor，该系统汇聚行业专家知识，结合 6 大类 50 余种社会经济发展综合决策模型，推动政府科学决策水平和决策能力现代化，助推“数据治国”。

2、数据资产安全管理

一是在数据分级分类方面，可以通过应用机器学习、模式聚类、自然语言处理、语义分析、图像识别等技术，提取数据文件核心信息，

对数据按照内容进行梳理，生成标注样本，经过反复的样本训练与模型修正，可以实现对数据自动、精准的分级分类。例如，我国网络安全初创企业思睿嘉得利用无监督机器学习引擎分析大量未经标注的原始文档集，自动按照内容进行主题梳理，并通过人工干预灵活调整语义相似度，获得满意的聚类效果，从而实现对数据的精准分级分类。浙江省旅游信息中心联合厦门杜若科技公司开展了浙江省旅游度假区信息的数据治理试点，将旅游大数据纳入人工智能系统，对结构化数据进行开放式训练，对数据进行分级分类并实现基于自然语言的数据管理。

二是数据质量管理方面，在开展数据质量核查过程中，人工智能技术与传统根据预置规则进行核查的方式相结合，可以仅针对少量核心核查规则，利用机器学习算法进行深度分析，定位数据质量原因、预测数据质量问题，形成知识库，进一步增强数据质量管理能力。例如，谷歌将人工智能引入医疗行业，通过重塑医疗数据层级为医疗巨头提供更高质量的结构化数据，创建新数据管道，助力医疗健康数据基础设施建设。亿信华辰的数据质量管理平台 EsDataClean，Informatica 的数据治理工具 Data Director 以及 IBM 的通用数据治理产品 Stewardship Center 等均在业界处于领先地位，通过人工智能技术的使用极大减少了人力投入和过程干预，提升了数据质量管理效率，也为后续的模型训练提供了更多高质量数据。

3、数据活动安全保护

一是数据本体安全保护方面，包括数据脱敏、数据防泄漏、数据

加密等。数据脱敏方面，在数据分级分类的基础上，结合数据合规性规则智能生成脱敏特征库，并与敏感数据识别智能关联，实现智能发现和自动脱敏，有效降低敏感数据泄露风险。亚马逊的智能识图工具 Rekognition 可以辅助医务人员进行医学图像脱敏。**数据防泄漏方面，**加州伯克利大学团队运用人工智能技术开发了一款手机 APP，能够自动扫描手机相册内的裸露照片，改为加设密码存在该 APP 中，并进一步从相册与云空间删除，彻底防止私密照片外泄。**数据加密方面，**谷歌大脑成功开发出两个独立的人工智能加密算法,不但能够防范第三方破解，而且还能够自我学习，破解其他人工智能加密算法。

二是数据活动网络安全保护方面，基于人工智能的网络安全防护手段相比传统基于静态规则的方法具有持续进化能力。新威胁的产生不断为训练集加入新的数据，通过人工智能算法和模型调优，可以快速查阅每个可疑文件数以百万计的特征，智能识别最轻微的代码冲突；对内外部网络流量中的元数据进行关联分析，实时检测异常流量；利用庞大的关联处理能力并行监测海量数据点，实时生成风险预测，发现并阻止设备或网络攻击。

恶意代码分析方面，中科院软件所提出基于文本分类技术的恶意代码检测工具“飞鼠”系统，能够对大量恶意代码样本进行及时、高效和准确检测，同时也具有一定的泛化能力，能够检测一定的未知样本。大连市公安局提出了基于人工智能技术的恶意代码变种检测技术，将恶意代码映射为图像,提取图像特征,建立人工智能模型,利用恶意代码家族图像样本集训练检测模型,能够快速识别恶意代码变种及其

家族,有效提高了检测效率和准确率。

边界安全防护方面，2018 年 11 月，华为发布业界首款智能防火墙，内置基于人工智能的高级威胁检测引擎，支持加密流量免解密威胁检测，通过联动云端为企业提供智能化的网络边界防护，威胁检出率达到 99%以上。2019 年 4 月，新华三集团发布人工智能防火墙业界新品，采用高性能的双 GPU 加双 CPU 的人工智能硬件架构，提供每秒万亿次的运算能力，结合数十种人工智能算法的软件开放平台，实现全面感知、深度学习和智能防护，改变了传统安全运维难、发现慢和响应差的状况。

入侵威胁监测方面，腾讯安全团队基于真实运行行为、系统层监控和人工智能芯片检测,利用神经网络算法和算法模型云端训练自主研发了腾讯 TRP-AI 反病毒引擎。该引擎具有抗免杀、高性能、实时防护、可检测 0Day 病毒等优势，可自动化训练，大大缩小了查杀周期和运营成本,可使病毒检测覆盖率达到 90%,检测准确率高达 99%。2017 年，IBM 发布用于网络安全领域的 “沃森”人工智能系统，能够提供云和端威胁的感知应对能力。

三是数据流转行为分析，通过自然语言处理、机器学习、聚类算法对采集的基础数据进行行为建模，多维度勾勒出用户行为特征，形成用户画像知识图谱，实现智能化用户行为分析。同样，通过人工智能技术也可以对数据传输行为进行智能统计和关联分析,绘制数据流转动态图谱，有利于跟踪敏感数据走向，分析数据安全态势。例如，荣之联推出的智慧商业情报大数据平台依托人工智能技术建立用户

行为数据计算模型和情感交换计算模型，通过用户行为数据流转分析来预测用户行为可能性。

四是数据安全风险评估，经过训练后的神经网络算法能够解决具有相似特点的风险评估问题，通过对风险因素的学习，可以自动实现从输入到输出的复杂映射关系，对优劣性受多种因素综合影响的事物作出合理的综合评价，从而减少传统专家评估过程中主观分数的片面性影响。例如思维世纪推出基于人工智能技术的数据安全评估解决方案，对数据全生命周期中各个环节的数据脱敏状态、应用通道、使用行为等因素进行智能关联分析，得出数据安全风险评估结果，并根据评估结果进一步优化数据安全管理策略。

4、数据安全事件管理

人工智能技术由于其普适性、自学习、高效性等特点能够在数据处理环节应对更加复杂的数据结构和数据环境，得出更加严谨和稳固的模型和推演结果，完成更自主的信息捕捉、更智慧的分析判断和更智能的服务。在数据安全事件管理中，利用人工智能技术对网络中的数据进行自动爬取和深度挖掘分析，能够提高网络中敏感数据、有害信息的自动发现和识别效率，实现数据安全事件智能监测和预警。结合用户行为画像和数据安全态势图谱，人工智能技术能够对数据安全事件的源头进行追溯，从而辅助管理部门采取相应措施实现快速处置，显著提升数据安全事件的管理水平。

不良信息治理方面，百度推出的“人工智能+广告打假”仅 2018 年上半年处理了 145.4 亿条有害信息，其中占比居前两位的是淫秽色

情类和赌博类，分别为 51.04%和 16.63%。2019 年阿里巴巴推出“人工智能谣言粉碎机”，通过分析用户画像、与知识图谱里的权威知识库作匹配验证等步骤实现对新闻内容的智能可信度识别，在特定场景中的准确率已达到 81%。中国信息通信研究院基于所积累的标准样本库，开展对淫秽色情、涉恐涉暴等违法信息识别的建模训练，初步实现基于人工智能技术的不良信息检测能力，识别准确率在 97%以上，比传统方式提升了 17%，识别速度是传统方式的 110 倍。2018 年 2 月，英国内政部宣布了一项新的智能内容识别工具，利用人工智能技术在线自动检测互联网平台上的恐怖分子宣传内容，精确度达到 99.995%。

互联网反欺诈方面，我国人工智能初创企业第四范式开发的“人工智能+金融”服务平台，构建了亿级别的高维机器学习模型，能够高效、精准识别欺诈交易，智能反洗钱。该平台在某银行线上 B2C 交易欺诈防控准确率达 83%，较传统专家规则方式提升 316% ，比专家规则多识别欺诈交易 58.8%，降低 30%的交易案宗审核成本。阿里自研的“钱盾”反诈预警系统，利用人工智能技术助力警方预警拦截诈骗事件，9 个月内劝阻 8.7 万人，止损 6.9 亿元。中国信息通信研究院使用人工智能技术多维度分析不同的可疑特征，有效实现了互联网诈骗行为的识别和预警，其中涉诈网站识别准确率达到 95%，涉诈账号识别准确率达到 90%，仿冒 APP 识别准确率达到 92%。

打击数据黑产方面，腾讯守护者计划基于长期积累的人工智能技术能力，引入多维度的动态验证机制对抗数据黑产。运用人工智能技术协助警方刑事打掉“快啊答题”、“光速打码”两个团伙，这两个团

伙是国内最大的利用人工智能破解识别验证码的打码黑产团伙。

总之，人工智能技术已在数据安全治理的细分领域开展诸多应用，但是人工智能技术并不是万能的，构建可管、可控、可信的数据安全治理技术支撑体系仍面临诸多挑战。欧洲市场研究和咨询服务公司 kbv research 2017 年发布市场研究预测报告指出，数据安全市场将每年以 18%的复合增长率发展，估计 2023 年将达到 209 亿美元；若以在 2023 年达到全球 20%的 GDP 来看，中国市场规模将达到大约 400 亿元人民币，未来人工智能在数据安全治理领域仍存在很大应用潜力。然而，同样要理性认识到，人工智能作为一项新兴的底层通用技术，并不是为某一项应用特制，因此并不能解决数据安全治理的所有难题。例如在数据运营活动的网络安全防护技术手段方面，人工智能技术并不适用于某些 APT 攻击的场景，有些 APT 攻击针对性强，攻击行为的成功往往是孤例，不足以支持海量攻击样本库生成，传统方式在此类场景仍然十分有效。数据安全治理是一个全球性的话题，除人工智能技术以外，网络环境安全防护能力的升级、数据安全治理政策和规则的制定等都影响数据安全治理的效果和能力。

四、 国内外人工智能数据安全治理动态

当前，世界主要国家均在人工智能发展战略、伦理规范方面提出人工智能数据安全相关规划和基本原则，但相关法律法规还不够细化完善，安全技术研究方兴未艾，安全标准也处于制定初步阶段，人工智能数据安全治理工作任重道远。

（一） 国内外人工智能数据安全战略规划情况

世界主要国家把发展人工智能作为提升国家竞争力、维护国家安全的重大战略，加紧出台规划和政策，力图在新一轮国际科技竞争中掌握主导权。在数据安全方面，各国结合本国实际国情和人工智能发展情况，在相关发展战略中形成有针对性的规划建议。

1、美国：推动训练数据集建设，加强数据安全风险应对

一是推进高质量训练数据集的建设与开放。2016 年 10 月，美国连续发布《为人工智能的未来做好准备》和《国家人工智能研究和发展战略规划》两份报告，提出实施“人工智能公开数据”计划，实现大量政府数据集的公开，增强高质量和完全可追溯的联邦数据、模型和计算资源的可访问性，并开发用于人工智能训练、测试的公共数据集。2019 年 2 月，美国总统特朗普签署《人工智能倡议》发展规划，进一步指示加强联邦政府、机构的数据、算法和计算机处理资源对人工智能研发人员和企业的开放。**二是加强对数据安全问题的应对**。2019 年 6 月，美国发布新版《国家人工智能研发与发展战略计划》，要求所有机构负责人审查各自联邦数据和模型，注重保护数据安全、隐私和机密性。

2、欧盟：细化人工智能数据规则，关注个人数据与权益保护

2018 年 3 月，欧洲政治战略中心发布《人工智能时代：确立以人为本的欧洲战略》，战略中认识到欧洲人工智能发展面临数据短缺和数据偏见等问题，提出扩大人工智能系统所需数据源，设计利于欧洲数据收集、使用和共享的监管方案，确保《通用数据保护条例》(GDPR)

个人数据保护要求实施的建议。2018 年 4 月，欧盟委员会发布《欧盟人工智能》发展战略，建议公共政策应鼓励更广泛地分享私人数据，并遵守关于个人数据保护的法律政策。为最大程度地促进数据流转和分享，欧盟委员会将修订公共部门信息开放指令，出台私营部门数据分享指南，修订科研信息获取和保存建议，以及出台医疗健康数字化转型政策。2018 年 12 月，为落实《欧盟人工智能战略》，欧盟发布《人工智能协调计划》，将提供更多数据、确保信任等作为关键领域发力，并提出必须遵从《通用数据保护条例》的关键原则。

3、英国：强化数据安全监管，规范数据资源开发利用

2016 年 11 月，英国政府科学办公室发布《人工智能：未来决策制定的机遇与影响》。报告指出，为了促进负责任的创新和获得公众的信任，同时为投资者和发明者创造一个好的环境以及为科技发展争取合理的数据使用，英国政府必须采用负责任的态度和积极应对的监管方式。2018 年 4 月，英国政府发布《产业战略：人工智能领域行动》，提出改进现有的数据基础设施：发布更高质量的公共数据，设立地理空间委员会以改进对地理空间数据的访问，为数据共享和使用提供法律保障等。在数据安全方面，提出开发公平、安全的数据共享框架：与公私部门的主要数据持有者及数据科学社区合作，确定数据共享障碍；与业界合作探索安全、公平的数据传输框架与机制。

4、日本：构建数据驱动与知识驱动融合型人工智能，鼓励协同开展数据安全与隐私保护技术研究

2018 年 4 月，日本发布第五版《下一代人工智能和机器人核心

技术开发计划》，进行下一代人工智能研发布局。计划提出，探索构建数据驱动与知识驱动融合型人工智能，将知识与数据相融合，辅助人类进行推理与决策；开展下一代人工智能框架与核心模块研究，研究兼顾数据安全与隐私保护的数据获取技术，探讨复杂问题和复杂场景下人工智能多模块融合效率与性能提升的方法。同时，加大从美国引进人工智能人才的力度，促进双方青年共同开展研究，在数据安全、隐私保护等方向培养下一代研究人员。

5、印度：充分挖掘本国人工智能发展优势，关注数据安全和隐私保护

2018 年 6 月，印度发布《人工智能国家战略》报告，指出印度人工智能发展的优势与问题，特别关注军事安全与道德隐私领域，并就印度人工智能国家战略的构建提出了框架方案。报告认为，印度人工智能发展的目标在于成为发展中国家的人工智能中心，基于成熟的软件行业，印度多元的文化环境将为推进人工智能发展带来意想不到的贡献。**关于数据偏差**，报告指出数据偏差导致的算法决策缺乏中立性，建议“识别内置偏差，评估其影响，并找到减少数据偏差的方法”。**关于数据保护**，报告建议建立数据保护框架和部门监管框架，并促进采用国际标准。**关于隐私保护**，报告呼吁“采取适当的措施来缓解隐私泄露风险，并强调使用人工智能情况下采取更高标准的隐私保护的重要性”。

6、我国：高度重视数据集建设，推进人工智能安全应用，防范人工智能数据风险

一是高度重视基础数据集建设，推进数据开放。2016 年，发改委发布《互联网+人工智能三年行动实施方案》提出加快建设文献、语音、图像、视频、地图等多种类数据的海量训练资源库和基础资源服务公共平台。2017 年 7 月，国务院印发《新一代人工智能发展规划》，指出“重点建设面向人工智能的公共数据资源库、标准测试数据集、云服务平台等”以及“完善落实数据开放与保护相关政策，开展公共数据开放利用改革试点，支持公众和企业充分挖掘公共数据的商业价值，促进人工智能应用创新”。2017 年 12 月，工信部发布《促进新一代人工智能产业发展三年行动计划(2018-2020 年)》提出“到 2020 年人工智能产业支撑体系基本建立，具备一定规模的高质量标注数据资源库、标准测试数据集建成并开放”以及“加强行业对接，推动行业合理开放数据”。**二是推进人工智能安全应用。**《新一代人工智能发展规划》提出，促进人工智能在公共安全领域的深度应用，推动构建公共安全智能化监测预警与控制体系。《行动计划》提出，推动人工智能先进技术在网络安全领域的深度应用，加快漏洞库、风险库、案例集等共享资源建设。**三是加强人工智能数据风险防范。**《新一代人工智能发展规划》在促进人工智能发展的同时，关注人工智能数据安全风险，提出“强化数据安全与隐私保护，为人工智能研发和广泛应用提供海量数据支撑”以及“促进人工智能行业和企业自律，切实加强管理，加大对数据滥用、侵犯个人隐私、违背道德伦理等行为的惩戒力度”。

综合看，我国人工智能发展战略对人工智能数据安全进行了整体

规划。但是与国外相比，我国在战略落地实施中存在如下问题：**一是**在数据集建设过程中，政府和行业数据开放力度不足，缺乏有影响力的公共数据集。**二是**在数据安全治理实践中，侧重人工智能在安全领域应用，人工智能数据安全风险防范的技术研究和手段建设相对滞后。

（二） 国内外人工智能数据安全伦理规范情况

国外先进国家较早重视人工智能数据安全伦理原则。在企业层面。谷歌提出的人工智能“七原则”包含隐私原则：给予通知和同意的机会，鼓励具有隐私保护的架构，并提供适当的透明度和对数据使用的控制。微软提出的人工智能“六原则”包含“隐私与保障”原则：在设计人工智能时，必须要考虑智能隐私保护，必须要有先进的、值得信赖的保护措施，确保个人和群体的隐私信息安全。**在行业层面，**2017 年 1 月，阿西洛马人工智能 23 原则形成并发布，霍金、马斯克等近四千名各界专家签署支持。关于隐私保护方面，相关原则要求人工智能系统分析使用数据时，人类应当拥有对其自身产生的数据的访问、管理以及控制的权利；并且人工智能基于个人数据的应用不能削减人们真实的或者感知上的自由。**在国家和地区联盟层面。**2018 年 4 月，英国议会发布《英国人工智能发展计划、能力与志向》，提出了“人工智能不应用于削弱个人、家庭乃至社区的数据权利或隐私”等 5 项人工智能基本道德准则。2019 年 4 月，欧盟委员会发布了《可信赖人工智能伦理指南》，指出人工智能系统必须确保隐私和数据保护，这既包括用户提供的信息，也包括用户在和系统交互过程中生成的信息，同时确保收集的数据不会用于非法地或不公平地歧视用户的行为。

我国近年来加强人工智能数据安全伦理研究与制定。在企业层面，2019 年 7 月，腾讯、旷视科技等企业相继发布人工智能伦理准则。腾讯人工智能伦理报告《智能时代的技术伦理观——重塑数字社会的信任》指出，人工智能技术伦理观包含技术信任、个体幸福和社会可持续三个层面。其中，个体幸福要求确保人人都有追求数字福祉、幸福工作的权利，在人机共生的智能社会实现个体更自由、智慧、幸福的发展。旷视科技《人工智能应用准则》明确提出，人工智能解决方案的开发及使用过程中，需严格保护用户的个人隐私、保障数据安全。**在行业层面，**2018 年 9 月，《人工智能安全发展上海倡议》在世界人工智能大会期间发布。倡议提出人工智能发展需要保障用户的数据安全，不得以牺牲用户隐私为代价，需要加强数据保护立法，丰富人工智能的技术路线，不断强化人工智能应用中的用户隐私保护。2019 年 5 月，《人工智能北京共识》发布，包含“实现人工智能系统的数据安全”、“避免数据与平台垄断”、“建立合理的数据与服务撤销机制”等内容。2019 年 6 月，中国人工智能产业发展联盟发布《人工智能行业自律公约（征求意见稿)》，“保护隐私”原则要求，坚持以合法、正当、必要的原则收集和使用个人信息，加强对未成年人等特殊数据主体的隐私保护，强化技术手段，确保数据安全。**在国家层面，**2019 年 6 月，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，将“尊重隐私”作为八项原则之一，要求人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权；在个人信息的收集、存储、处理、使用

等各环节应设置边界，建立规范；完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为。

可以看到，伴随人工智能技术和应用发展，我国日益重视人工智能伦理规范研究，国家、行业和企业层面均已形成人工智能数据安全伦理规范。但是，由于相关伦理原则大多为近期发布，加之长期以来社会公众对个人数据保护的意识观念和重视程度存在较大差异，导致人工智能数据安全伦理的社会影响力受限，尚未真正形成社会共识。

（三） 国内外人工智能数据安全法律制定情况

世界主要国家的数据管理和隐私保护法案促进人工智能行业健康发展。一是数据安全要求得到明确细化，指导人工智能行业合规实践。2018 年 5 月，欧盟《通用数据保护条例》（GDPR）颁布，建立了用户个人信息访问、修正和删除请求相关机制，赋予欧盟用户控制个人数据的权力，成为各国制定个人信息保护法案的重要参考。其后印度《2018 年个人数据保护法案（草案）》、巴西《通用数据保护法》、美国《2018 年加州消费者隐私法案》等纷纷效仿《通用数据保护条例》（GDPR），对数据处理者的个人数据的收集和使用行为加以明确规范，促使人工智能行业进一步规范数据收集和使用行为。**二是**努力平衡数据权利保护与数据开放流动，促进人工智能发展。2018 年 11 月，欧盟通过《非个人数据在欧盟境内自由流动框架条例》，致力于为企业和公共部门清除欧盟内部非个人数据自由流动障碍。2018 年 12 月，美国国会通过《开放政府数据法案》，要求联邦机构必须以“机器可读”和开放的格式发布任何“非敏感”的政府数据并使用开放许可协

议。**三是**为人工智能数据安全监管提供了法律依据。各国个人信息保护法案对企业违规列出明确处罚规定，可作为行政部门进行人工智能数据安全监管的有力依据，对相关企业产生威慑效应。2019 年 1 月，法国数据保护机构（CNIL）依据《通用数据保护条例》（GDPR）对谷歌开出 5000 万欧元罚单。2019 年 7 月，美国联邦贸易委员会（FTC）对“剑桥分析事件”的 Facebook 处以 50 亿美元罚款。**四是**相关法案通过数据匿名化加强人工智能数据保护。为了避免个人数据被挖掘和滥用，欧盟《通用数据保护条例》（GDPR）、日本《个人信息保护法》等通过数据匿名化的方式进行个人数据保护。数据匿名化能够降低人工智能数据泄露风险，有利于人工智能行业健康发展。

我国立足现有法律基础，加速完善数据安全保护立法。一是我国现行法律涉及人工智能数据安全相关内容，具备一定法律基础。**在国家法律层面**，2009 年，《刑法修正案（七）》首次将特定主体的个人信息保护义务与责任写入刑法，规定了出售、非法提供公民个人信息罪和非法获取公民个人信息罪。《刑法修正案（九）》根据打击个人信息犯罪的实际需要，将罪名调整为“侵犯公民个人信息罪”。2012 年，《全国人民代表大会常务委员会关于加强网络信息保护的决定》明确了网络服务提供者的义务和责任，并赋予政府主管部门必要的监管手段，以保护公民个人信息安全。2016 年，《网络安全法》增加了最少够用原则、信息权利人删除权、知情权、更正权等新规定，实现了与国际规则和欧美个人信息保护立法理念接轨。2019 年 1 月，《电子商务法》正式实施，提出个人信息收集和使用保存的最小化、将用户知

情强化为用户明示同意等细化要求。现有法律为人工智能领域数据安全保护提供了基本依据。**在部门规章层面**，为落实国家法律法规管理要求，政府部门重点针对个人信息保护出台相应管理文件，保护个人信息安全和个人信息主体合法权益。工业和信息化部《电信和互联网用户个人信息保护规定》进一步明确电信业务经营者、互联网信息服务提供者收集、使用个人信息的规则和信息安全保障措施。公安部《公安机关互联网安全监督检查规定》明确规定了互联网安全监督检查过程中的个人信息保护要求和处罚措施等内容。**二是**加速推进数据安全保护立法和人工智能专门立法。我国目前正在多层面推进数据安全和个人信息保护法律法规等规范制定，加速完善相关保护和监管规则，既包括国家层面的基本立法，如《数据安全法》、《个人信息保护法》，也包括部委层面的规章和规范性文件，例如最近公开征求意见的《数据安全管理办法》、《个人信息出境安全评估办法》等。同时，全国人大常委会表示，已把人工智能方面立法列入抓紧研究项目，努力为人工智能的创新发展提供有力的法治保障。

整体来看，我国目前尚未形成体系完善的人工智能数据安全法律法规，并且，数据安全和个人信息保护立法相对滞后，相关规定散落在《民法总则》、《网络安全法》、《电子商务法》等法律法规中，亟需针对当前新技术和新业态发展，加速完成数据安全和个人信息保护的顶层立法，为人工智能健康发展提供法律支撑。

（四） 国内外人工智能数据安全技术发展情况

主要国家积极推进人工智能数据安全技术研究。一是加大人工智

能数据安全相关研究资金投入。2018 年 9 月，美国国防高级研究计划局（DARPA）投资 20 亿美元启动 AI Next 项目，致力于开发第三代人工智能技术。其中，包括对抗性人工智能和高性能人工智能等。对抗性人工智能防止输入异常数据造成的智能系统运行错误；高性能人工智能可降低算法对训练数据的强依赖性。**二是**加强人工智能数据安全基础理论研究。重点研究减少训练数据量的人工智能基础理论方法。迁移学习研究将已训练好的模型参数进行迁移，来提升新模型训练效率，使人工智能系统在不收集大量原始数据的情况下解决新问题，通过减少数据需求量来降低数据安全风险。联邦学习研究在客户端利用本地数据进行分布式训练，从而数据不用上传至服务器，在不泄露用户个人数据的情况下更新人工智能算法模型，有效保证数据安全性。谷歌 Gboard 利用联邦学习，基于分散独立设备的数据、词汇大大提升了推荐准确性。**三是**积极攻克人工智能数据安全关键技术。重点加强人工智能数据加密技术研究。差分隐私研究提升人工智能系统的用户隐私保护能力，使人工智能系统数据集包含噪声，确保特定用户个人隐私的机密性。苹果公司将差分隐私技术应用于智能终端产品，用以保护用户隐私信息。同态加密研究使人工智能系统直接使用加密后的数据训练模型，且不会影响模型的有效性和可用性。

我国研究机构和企业同步开展人工智能数据安全技术研究，在部分领域取得较好进展。2018 年 7 月，清华大学创业公司瑞莱智慧成立，研究实现减少标注数据数量、决策可解释、模型安全可靠相关技术，团队近年来开发的“珠算(ZhuSuan)”概率编程库，可减少实际

场景中需要的标注数量。第四范式公司作为国内迁移学习实践领跑者，已将迁移学习算法应用到公司核心产品“先知”平台，并在医疗领域实现落地应用。2019 年 6 月，微众银行人工智能团队开源全球首个工业级的联邦学习框架 FATE，并将相关成果贡献给 Linux 基金会，加强了我国在人工智能数据安全领域的行业地位，加快联邦学习技术在数据安全方面的落地进程。

由上可见，针对人工智能数据安全风险，相关技术研究正处于起步阶段。美国等西方国家凭借先发技术优势，加大研发投入，提升人工智能安全能力。我国作为数字经济大国和人工智能先行国家，需从国家层面加强规划引领和资金投入，维护数据安全，保障基于信任和安全的数据流动，促进人工智能数据安全技术研究应用。

（五） 国内外人工智能数据安全标准规范情况

国际标准化组织积极研究人工智能数据安全相关标准，ISO/IEC JTC1 SC42 WG3 人工智能可信标准组正在开展人工智能风险管理、人工智能的可信度概览等标准研制。IEEE 标准协会对涉及人工智能道德规范的伦理标准进行研究，包括 P7002 数据隐私处理、P7004 儿童和学生数据治理标准、P7005 透明雇主数据治理标准与 P7006 个人数据人工智能代理标准等。IEEE P3652.1 联邦学习基础框架与应用工作组已开展联邦学习的相关标准化工作。**区域和国家标准化组织**开始重视人工智能数据安全标准，2019 年 5 月 1 日，美国国家标准与技术研究院（NIST）发布人工智能标准化计划纲要，将人工智能数据安全与隐私保护相关标准化纳入人工智能可信标准领域。

我国全国信息安全标准化技术委员会（SAC/TC260）、中国通信标准化协会（CCSA）等标准化组织积极推进人工智能数据安全相关标准制定工作。**TC260 在生物识别、智能终端、大数据、个人信息保护等领域开展了数据安全相关标准化工作。**在生物识别领域，开展了《信息安全技术 指纹识别系统技术要求》与《信息安全技术 虹膜识别系统技术要求》标准研制，对生物识别系统的数据保护能力提出要求；在移动智能终端领域，开展了《信息安全技术 移动智能终端个人信息保护技术要求》标准研制，对移动智能终端中的个人信息与数据保护能力提出要求；在大数据领域，开展了《信息安全技术 大数据服务安全能力要求》标准研制，对人工智能相关的大数据安全能力提出要求。在个人信息保护领域，开展了《信息安全技术 个人信息安全规范》标准研制，明确了个人信息的收集、保存、使用、共享的合规要求，为人工智能行业数据安全和隐私保护提供重要参考。在人工智能安全领域，中国电子技术标准化研究院牵头开展人工智能安全标准框架研究以及《信息安全技术 人工智能算法安全指南》标准研制，将人工智能数据安全列为重要研究内容。**CCSA 在生物识别、人工智能终端、人工智能服务平台、数据安全保护等领域开展了数据安全相关标准化工作。**在人工智能终端领域，开展《人工智能终端产品 个人信息保护要求和评估方法》与《人工智能终端设备安全环境技术要求》标准研制，对人工智能终端的个人信息保护与终端设备环境的安全能力提出要求。在人工智能服务平台领域，开展《人工智能服务平台数据安全要求》标准研制，对人工智能服务端的数据安全管理与评

估提出要求。在数据安全保护领域，成立数据安全特设组，整合资源对数据分级分类、数据安全合规性要求等重要标准进行研究制定。

目前，国内外人工智能数据安全以及隐私保护标准大都处于制定阶段，我国在《数据安全法》和《个人信息保护法》尚未出台的情况下，相关标准起到了行业指引作用，得到业界重视。但是，人工智能安全标准体系尚未形成，人工智能数据安全收集、使用和共享等关键技术标准尚未形成，亟需构建人工智能数据安全标准体系和发展规划，并加快制定实施。

五、 人工智能数据安全治理建议

当前，人工智能处于技术发展和应用普及快速迭代时期，人工智能数据安全风险不断凸显，安全应用逐步深化，问题挑战与发展机遇相伴而生。我国作为数字经济大国和人工智能先行国家，需坚持发展与安全并重的治理思路，以伦理规范为引导，以法律法规为底线，以安全监管为约束，大力推进标准建设、技术发展和人才培养等工作，全面提升我国人工智能数据安全的综合治理能力，有效保障我国数字经济和智能社会的健康稳步发展，维护人民利益和国家安全，确保人工智能数据安全、可靠、可控。

（一） 明晰发展与安全并举的治理思路

一是推进人工智能数据资源建设，在发展中解决安全问题。建立健全适合我国国情的数据流通共享机制，推动政府和行业数据开放，培育规范数据交易市场，鼓励不同市场主体安全的进行数据交换，构建支撑我国人工智能产业发展的优质数据资源，在发展中规避数据偏

见、数据权属等人工智能数据安全问题。**二是加强人工智能数据安全治理能力，以安全促进发展。**基于人工智能数据安全风险研究，依托现有数据安全管理机制和技术手段，加大人工智能应用场景下数据安全防护技术研究，同时，促进人工智能技术在数据安全治理与网络攻防对抗等领域中的应用，实现人工智能数据安全风险的提前感知和预防，规避训练数据污染、数据智能窃取等数据安全风险，促进人工智能安全发展。

（二） 引导社会遵循人工智能伦理规范

一是加强人工智能伦理原则的社会宣贯。针对我国人工智能治理机构、行业和企业发布的人工智能伦理原则，加强社会宣传教育，加大社会影响范围，真正形成社会共识，使其成为人工智能参与方在设计、研发、使用、治理过程中的潜在道德观念，提升人工智能用户人群特别是青少年的个人数据和权益保护意识，降低人工智能发展过程中可能存在的数据安全伦理风险。**二是积极参与国际人工智能伦理规范制定。**通过联合国、G20、亚太经合组织、上合组织等国际平台，积极开展国际对话与合作，在充分尊重各国人工智能治理原则和实践的前提下，贡献我国人工智能数据安全治理思路，推动形成具有广泛共识的国际人工智能数据安全伦理规范。

（三） 建立人工智能数据安全法律法规

一是推进人工智能和数据安全相关立法工作。在国家层面，推进《数据安全法》、《个人信息保护法》以及人工智能相关法律出台，明确人工智能数据安全法律原则，确立不同参与主体在人工智能生命周

期各阶段所享有的数据权利与承担的安全责任，设立人工智能数据安全问责制和救济制度，并对人工智能相关数据过度采集、偏见歧视、资源滥用、深度伪造等突出问题进行规制，为人工智能数据安全管理提供基本法律依据。**二是完善人工智能数据安全相关部门规章。**依据国家相关法律，结合人工智能在不同领域应用中的特点，针对各领域关键突出人工智能数据安全风险，制定和细化相关部门规章，提出对所属领域的人工智能算法设计、产品开发和成果应用等过程中数据安全要求。**三是开展人工智能数据安全执法。**加强对人工智能数据收集、使用、共享等高风险环节安全执法，特别是对数据过度采集、数据资源滥用、侵犯个人隐私、违背道德伦理等行为加大执法惩戒力度，创新和规范人工智能数据安全事件调查取证方法和程序，促进人工智能数据安全法律和规章有效落地执行。积累执法经验并总结不足，形成反馈机制持续完善相关法律和部门规章。

（四） 完善人工智能数据安全监管措施

一是开展人工智能数据安全监督惩戒。依照国家法律法规，政府部门针对数据过度采集、数据偏见歧视、数据资源滥用等人工智能数据安全风险，通过线上线下多种方式实施监督检查，及时发现和防范安全隐患。针对基于人工智能的网络攻击、深度伪造等严重不良行为，利用技术手段监测和社会公众监督等方式，及早发现，降低危害，加强惩戒。**二是开展人工智能数据安全检测评估。**依托行业组织或者第三方机构，构建人工智能数据安全检测评估平台，制定人工智能产品、应用和服务的数据安全检测评估方法和指标体系，研发安全检测评估

工具集，通过测试验证提升人工智能产品安全性和成熟度，降低人工智能数据安全风险。通过检测评估强化企业的数据安全与隐私保护，为人工智能研发和广泛应用提供海量数据支撑。

（五） 健全人工智能数据安全标准体系

一是完善我国人工智能数据安全标准体系，加快急需重点标准研制。在我国人工智能安全标准框架下，加快研制人工智能数据安全标准体系，制定人工智能数据安全标准推进计划。重点加快推进人工智能数据安全评估、人工智能平台数据安全保护、自动驾驶用户隐私保护等行业急需重点标准研制工作。**二是优化我国人工智能数据安全标准化组织建设。**推动国家信息安全标准化技术委员会、中国通信标准化协会等国家及行业标准化组织成立人工智能安全研究组，促进国家、行业和团体标准化组织联合有序推进人工智能数据安全标准出台。**三是加强国际人工智能数据安全标准化工作。**组织国内企业、科研院所等多方力量加强研究储备，在IEEE、ISO/IEC、ITU等国际标准化组织中联合发声，提出更多人工智能数据安全相关提案，贡献更多中国力量和方案，实质性参与和主导人工智能数据安全相关国际标准工作。

（六） 创新人工智能数据安全技术手段

一是加强人工智能数据安全保护基础理论研究和技术研发。利用国家专项和社会基金引导产学研各界联合开展人工智能数据安全风险产生机理和防御理论的研究，并突破小样本学习、联邦学习、差分隐私等人工智能数据安全保护核心关键技术。**二是建设完善我国人工智能开源学习框架，提供保障数据安全的人工智能基础研发平台。**鼓

励企业建设完善人工智能开源学习框架，增强框架内置数据安全设计和技术措施。并且通过我国市场优势，加快培育自有人工智能开源平台共享应用生态圈和产业链。**三是促进人工智能在数据安全领域中的应用。**鼓励人工智能企业和数据安全企业充分发挥各自优势，通过成立联合实验室、共同投资等多种方式，开展人工智能技术在数据安全治理领域的应用研究和产品技术研发。

（七） 培养复合人工智能数据安全人才

一是完善学校人工智能数据安全教育。鼓励高校尽快形成人工智能与网络信息安全交叉学科的人才培养模式，组建和壮大人工智能安全师资队伍，促进国内外人工智能安全学生和教师共同开展研究，扩大人工智能数据安全人才培养规模、提高人工智能数据安全人才培养质量。**二是加大企业人工智能数据安全人才培养。**鼓励企业内部创办培训机构，或与科研机构、高校等建立联合人工智能数据安全培训基地，加强企业人员人工智能数据安全管理和技术能力培训。**三是加强国外人工智能数据安全人才引进。**制定人才政策引进专项人才，支持高校或企业引进世界一流人工智能数据安全领军人才；鼓励企业通过资本运作等方式吸纳掌握核心技术的人工智能数据安全团队。

致　谢

本白皮书在撰写过程中得到了中国信息通信研究院政策与经济研究所、泰尔终端实验室以及深圳市腾讯计算机系统有限公司、阿里巴巴（中国）有限公司、北京字节跳动科技有限公司、网易（杭州）网络有限公司等单位的大力支持，特此感谢！

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62305900

传真：010-62300264

网址：www.caict.ac.cn